多元随机变量

司继春

[1] 上海对外经贸大学

2024年9月

概览

多元随机变量

多元随机变量/随机向量的定义

给定一个概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，一个$n$维的随机向量$X$即从样本空间到$n$维欧几里得空间的函数，$X : \Omega \to \mathbb{R}^n$。

向量表达

注意以上定义我们使用了向量的表达方式，即：

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

多元随机变量

4	5	6	7	8
3	4	5	6	7
2	3	4	5	6
1	2	3	4	5
	1	2	3	4

Figure: 四面骰子

多元随机变量

多元随机变量

投两个均匀的四面骰子，则$\Omega = \{(1,1),(1,2),\ldots(4,4)\}$:

- 定义随机变量X_1为两个骰子的数值之和
- 定义X_2为两个骰子中较小的骰子的数值，如上图所示。

那么向量$X = [X_1, X_2]' : \Omega \to \mathbb{R}^2$为一个随机向量，其可能的取值为

$$\left\{[x_1, x_2]', x_1 \in \{2, ..., 8\}, x_2 \in \{1, 2, 3, 4\}\right\}$$

例如，$X^{-1}\left(\{[5,2]'\}\right) = \{(2,3),(3,2)\}$

随机向量的概率

进而，我们可以使用$(\Omega, \mathscr{F}, \mathscr{P})$和一个随机向量$X$的定义导出一个$(\mathbb{R}^n, \mathscr{B}^n)$上的概率函数的定义。定义

$$P_X(A) = \mathscr{P}\left(X^{-1}(A)\right), \forall A \in \mathscr{B}^n$$

随机向量的概率

在四面骰子的例子中，如果$A = \{[5,2]'\}$，那么

$$P_X(A) = \mathscr{P}\left(X^{-1}(A)\right) = \mathscr{P}\left(\{(2,3),(3,2)\}\right) = \frac{2}{16}$$

同理，$P_X(\{[2,1]'\}) = \frac{1}{16}$，$P_X(\{[5,a]', a \in \{1,2,3,4\}\}) = \frac{4}{16}$等等。

联合分布函数

联合分布函数

由$(\Omega, \mathscr{F}, \mathscr{P})$导出的概率空间$(\mathbb{R}^n, \mathscr{B}^n, P)$的联合分布函数（joint c.d.f.）定义为：

$$F(x) = F(x_1, x_2, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n), \forall x \in \mathbb{R}^n$$

联合分布函数为单调递增且$F(-\infty, -\infty, ..., -\infty) = 0$，$F(\infty, \infty, ..., \infty) = 1$

联合密度函数

联合密度函数

1. 如果随机向量X的每个分量都是离散型随机变量，那么可以定义联合概率质量函数p.m.f为：

$$f\left(x_1, x_2, \ldots, x_n\right) = P\left(\left\{X_1 = x_1, \ldots, X_n = x_n\right\}\right)$$

2. 如果随机变量X的联合分布函数连续，如果函数$f\left(x\right)$满足：$P\left(X \in A\right) = \int_A f\left(x\right) dx, x \in \mathbb{R}^n, A \subset \mathscr{B}^n$那么我们称$f\left(x\right)$为其联合概率密度函数p.d.f。特别的，如果联合分布函数$F\left(x\right)$可微那么：

$$f\left(x_1, x_2, \ldots, x_n\right) = \frac{\partial^n F\left(x_1, x_2, \ldots, x_n\right)}{\partial x_1 \partial x_2 \cdots \partial x_n}$$

概率质量函数

概率质量函数

四面骰子例子中的概率质量函数可以用下表描述：

$X_2 \backslash X_1$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$

概率密度函数

概率密度函数

如果随机向量$X=[X_1,X_2]'$的两个分量分别服从正态分布，且相互独立，那么其概率密度函数为：

$$f(x)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}$$

边缘分布

- 既然随机向量的每个分量X_i都是一个随机变量，那么自然也存在分布函数$F_{X_i}(x_i)$。
- 如果已知联合分布函数，根据（联合）分布函数的定义，X_i的分布函数可以通过

$$F_{X_i}(x_i) = P(X_i \le x_i) = P(X_1 \le \infty, ..., X_i \le x_i, ..., X_n \le \infty) = F_X(\infty, ..., x_i, ..., \infty)$$

计算。
- $F_{X_i}(x_i)$即随机变量X_i的分布函数，在这里由联合分布函数计算出来，所以我们也将其称为联合分布函数$F(x)$的边缘分布函数（marginal c.d.f.），而对应的密度（质量）函数可以相应定义。

边缘分布

- 进一步拓展，如果$X=[X_1,...,X_n]'$为随机向量，那么其i_k个分量$\widetilde{X}=[X_{i_1},X_{i_2},...,X_{i_k}]',1\le i_1<i_2<...<i_k\le n$也是一个随机向量。
- $\widetilde{X}$的联合分布函数同样可以通过$F(x)$来定义，即令$F(x)$中满足$j\notin\{i_1,...i_k\}$的分量为∞ 。
- 比如，对于三维随机变量$X=[X_1,X_2,X_3]'$，则$\widetilde{X}=[X_1,X_2]'$的分布函数为：$F_{\widetilde{X}}(\widetilde{x})=F(\widetilde{x}_1,\widetilde{x}_2,\infty)$。
- 而对应的密度（质量）函数可以通过

$$f_{\widetilde{X}}(x)=\frac{\partial^2F(x_1,x_2,\infty)}{\partial x_1\partial x_2}=\int_{\mathbb{R}}\frac{\partial^3F(x_1,x_2,x_3)}{\partial x_1\partial x_2\partial x_3}dx_3=\int_{\mathbb{R}}f(x_1,x_2,x_3)\,dx_3$$

来计算。

边缘分布

边缘质量函数

四面骰子例子中，$X = [X_1, X_2]'$，X_1和X_2的边缘概率质量函数如下表所示：

$X_2 \backslash X_1$	2	3	4	5	6	7	8	F_{X_2}	f_{X_2}
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_{X_1}	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_{X_2}$ $\parallel$
f_{X_1}	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_{X_1} =$	1

边缘分布

边缘密度函数

上例中的联合正态分布，其边缘分布函数为:

$$\begin{aligned} F_{X_1}(t) &= \int_{\mathbb{R}} \int_{-\infty}^{t} f(x_1, x_2)\, dx_1 dx_2 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \int_{-\infty}^{t} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\} dx_1 dx_2 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \exp\left\{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\} dx_2 \int_{-\infty}^{t} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\} dx_1 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{t} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\} dx_1 \end{aligned}$$

则其边缘密度函数为:

$$f_{X_1}(t) = \frac{dF_{X_1}(t)}{dt} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(t-\mu_1)^2}{2\sigma_1^2}\right\}$$

边缘分布

注意如果只确定了边缘分布，联合分布并不能唯一确定。

联合分布与边缘分布

以下两个联合质量函数具有相同的边缘分布，然而其联合质量函数并不相同：

$X_2 \backslash X_1$	0	1	f_{X_2}
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
f_{X_1}	$\frac{1}{2}$	$\frac{1}{2}$	1

$X_2 \backslash X_1$	0	1	f_{X_2}
0	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{2}$
1	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
f_{X_1}	$\frac{1}{2}$	$\frac{1}{2}$	1

边缘分布与联合分布

联合分布与边缘分布

如果随机向量$[U,V]'$，其分布函数为：

$$F_{U,V}(u,v)=\begin{cases}1 & u,v>1\\ 0 & u<0\ or\ v<0\\ \min\{u,v\} & otherwise\end{cases}$$

其边缘分布：

$$F_U(u)=F_{U,V}(u,\infty)=\mathbb{1}\{0\le u\le 1\}\cdot u$$
$$F_V(v)=F_{U,V}(\infty,v)=\mathbb{1}\{0\le v\le 1\}\cdot v$$

即其边缘分布为均匀分布。

边缘分布与联合分布

联合分布与边缘分布

如果另一分布函数为:

$$\tilde{F}_{U,V}(u,v) = \begin{cases} 1 & u,v > 1 \\ 0 & u < 0 \; or \; v < 0 \\ \min\{u,1\} \cdot \min\{1,v\} & otherwise \end{cases}$$

其边缘分布也为均匀分布。

边缘分布与联合分布

- 以上示例中，虽然$F_{U,V}(\cdot,\cdot)$和$\tilde{F}_{U,V}(u,v)$的边缘分布函数相同，但是联合分布函数确实不相同的，那么意味着，如果我们仅仅知道$U\sim \mathrm{U}(0,1),V\sim \mathrm{U}(0,1)$，是无法反推回其联合分布函数的。
- 这两者的差别就在于：边缘分布仅仅建模了单个随机变量的分布，然而联合分布还包含了两个随机变量之间相依性（dependency）的信息。

多元随机变量的期望

- 与一元随机变量类似，对于随机向量X的数学期望可以使用Riemann-Stieltjes积分进行定义，不过我们在这里需要多元函数的Riemann-Stieltjes积分，其定义过程可以在一元的Riemann-Stieltjes积分基础上进行拓展，需要做一些特殊的定义。
- 一个难点是如何定义$\Delta F(x)$?
- 按照一元函数的定义，即$\Delta F(x) = F(x_2) - F(x_1)$显然是不合适的，因为这里$x_1, x_2$均为$n$维向量。
- 为此，我们需要定义在一个n维矩形

$$A = \times_{j=1}^{n} (a_j, b_j]$$

上的$\Delta F(x)$。

$\mathbb{R}^n$上的Riemann-Stieltjes积分

- 考虑A的2^n个端点

$$\mathscr{V}=\left\{v_1,\ldots,v_{2^n}\middle|v_m=\left[x_{1m},\ldots,x_{nm}\right]',x_{jm}\in\left\{a_j,b_j\right\}\right\}$$

其中如果$[x_{1m},\ldots,x_{nm}]'$中a_j的数量为偶数则$\text{Sign}\left(v_m\right)=1$，否则$\text{Sign}\left(v_m\right)=-1$。

- 使用如上记号，对于n维矩形A，$\Delta F\left(x\right)$可以被定义为

$$\Delta F_A=\sum_{v\in\mathscr{V}}\text{Sign}\left(v\right)F\left(v\right)$$

- 可以证明，如果$F\left(x\right)$可微，那么

$$\Delta F_A=\int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}f\left(x\right)dx_n\cdots dx_1$$

其中$f\left(x\right)=\frac{\partial^n F(x)}{\partial x_1\partial x_2\cdots\partial x_n}$。

$\mathbb{R}^n$上的Riemann-Stieltjes积分

$\Delta F(x)$的定义

对于二维平面上的矩形$A=(0,1]\times(2,3]$

- 四个顶点为$\{(0,2),(0,3),(1,2),(1,3)\}$，其中$(0,2)$和$(1,3)$分别包含了2个和0个左端点，从而$\text{Sign}((0,2))=\text{Sign}((1,3))=1$
- 而$(0,3)$和$(1,2)$都只包含1个左端点，从而$\text{Sign}((0,3))=\text{Sign}((1,2))=-1$。
- 从而

$$\Delta_A F=F(1,3)+F(0,2)-F(0,3)-F(1,2)$$

$\mathbb{R}^n$上的Riemann-Stieltjes积分

- 接下来，我们可以首先仿照一维上的定义，对每一维都做一个划分：$\Pi_j : a_j = x_{i0} < x_{i1} < \cdots < x_{i,m_j} = b_j$，那么积分区域$A$可以被划分为

$$A = \bigcup_{\mathscr{I}} \left(\times_{j=1}^{n} A_{j,i_j} \right)$$

其中$\mathscr{I} = \{(i_1, i_2, ..., i_n) \,|1 \le i_j \le m_j\}$，$A_{j,i_j} = \left(a_{j,i_j}, b_{j,i_j}\right]$。

- 如此，我们将$\mathbb{R}^n$中的一个“矩形”A分解为了很多小的矩形$R_J = \times_{j=1}^{n} A_{j,i_j}, J \in \mathscr{I}$。
- 进一步，定义

$$I = \sum_{J \in \mathscr{I}} g(x_J) \, \Delta F_{R_J}$$

其中x_J为矩形R_J中的某个值。

- 现在，令$mesh\left(\Pi_j\right) \to 0, j = 1, ..., n$，如果以上级数极限存在，就可以定义Riemann-Stieltjes积分$\int_A g(x) \, dF(x)$了。

$\mathbb{R}^n$上的Riemann-Stieltjes积分

- 如此，对于一个随机向量$X = [X_1, ..., X_n]'$及其联合分布函数$F(x)$，对于任意的函数$g(x): \mathbb{R}^n \to \mathbb{R}$，可以定义期望

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x)\, dF(x)$$

- 特别的，如果存在密度函数$f(x)$，那么如上期望可以使用

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) f(x)\, dx$$

计算。

- 根据此定义，如果令$g(X) = \iota_i' X = X_i$，其中$\iota_i = [0, 0, ..., 1, ...0]'$，有

$$\mathbb{E}(\iota_i' X) = \int_{\mathbb{R}^n} x_i dF(x) = \int_{\mathbb{R}} x_i dF_{X_i}(x_i) = \mathbb{E}(X_i)$$

即多元随机变量的分量的期望与一元随机变量的期望定义相同。

随机向量的期望

- 我们通常把随机向量的期望写为向量形式:

$$\mathbb{E}(X)=\begin{bmatrix}\mathbb{E}(X_1)\\ \mathbb{E}(X_2)\\ \vdots\\ \mathbb{E}(X_n)\end{bmatrix}$$

- 在这里期望的线性性仍然成立，比如，如果$\iota=[1,1,...,1]'$为全部由1构成的向量，那么:

$$\mathbb{E}(\iota' X)=\mathbb{E}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathbb{E}(X_i)$$

- 同理如果令$\mu=\mathbb{E}(X)=[\mathbb{E}(X_1),\mathbb{E}(X_2),...,\mathbb{E}(X_d)]'$，令$a\in\mathbb{R}^n$，那么我们有

$$\mathbb{E}(a'X)=\mathbb{E}\left(\sum_{i=1}^{n}a_iX_i\right)=a'\mathbb{E}(X)=a'\mu$$

随机向量的期望

随机向量期望的线性性

对于一个实数矩阵$A_{h\times n}=[a_1,a_2,...,a_h]'$，以及$h$维向量$b=[b_1,...,b_h]'$，有

$$\mathbb{E}\left(AX+b\right)=A\mathbb{E}\left(X\right)+b$$

Proof.

矩阵乘积为$AX=[a_1'X,a_2'X,...,a_h'X]'$，其期望为

$$\mathbb{E}\left(AX+b\right)=\mathbb{E}\left(\begin{bmatrix} a_1'X+b_1 \\ a_2'X+b_2 \\ \vdots \\ a_h'X+b_h \end{bmatrix}\right)=\begin{bmatrix} \mathbb{E}\left(a_1'X\right)+b_1 \\ \mathbb{E}\left(a_2'X\right)+b_2 \\ \vdots \\ \mathbb{E}\left(a_h'X\right)+b_h \end{bmatrix}=\begin{bmatrix} a_1'\mathbb{E}\left(X\right) \\ a_2'\mathbb{E}\left(X\right) \\ \vdots \\ a_h'\mathbb{E}\left(X\right) \end{bmatrix}+b=A\mathbb{E}\left(X\right)+b$$

$\square$

随机向量的矩母函数

矩母函数

对于随机向量$X=[X_1,...,X_n]'$，令$t=[t_1,...,t_n]'$，矩母函数可以定义为

$$M_X(t)=\mathbb{E}\left(e^{t'X}\right)$$

根据以上定义，如果令$\tilde{t}=[0,...,t_i,...,0]$，即$\tilde{t}$的第$i$个分量为$t_i$，其他分量为0，那么

$$M_X(\tilde{t})=\mathbb{E}\left(e^{\tilde{t}'X}\right)=\mathbb{E}\left(e^{t_iX_i}\right)=M_{X_i}(t_i)$$

即随机变量X_i的矩母函数。

协方差

对于随机向量$X=[X_1,X_2]'$，如果$\mathbb{E}\left(X_1^2\right)<\infty,\mathbb{E}\left(X_2^2\right)<\infty$，根据Cauchy-Schwarz不等式，

$$\mathbb{E}\left|X_1X_2\right|\leq\sqrt{\mathbb{E}\left|X_1\right|^2\mathbb{E}\left|X_2\right|^2}<\infty$$

即X_1X_2可积，我们可以定义两个随机变量的协方差（**covariance**）：

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(X_1,X_2\right)&=\mathbb{E}\left[\left(X_1-\mathbb{E}\left(X_1\right)\right)\left(X_2-\mathbb{E}\left(X_2\right)\right)\right]\\&=\mathbb{E}\left[X_1X_2-\mathbb{E}\left(X_1\right)X_2-X_2\mathbb{E}\left(X_1\right)+\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\right]\\&=\mathbb{E}\left(X_1X_2\right)-2\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)+\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\\&=\mathbb{E}\left(X_1X_2\right)-\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\end{aligned}$$

当$X_2=X_1$时，

$$\mathbb{C}\mathrm{ov}\left(X_1,X_1\right)=\mathbb{E}\left(X_1^2\right)-\left[\mathbb{E}\left(X_1\right)\right]^2=\mathbb{V}\left(X_1\right)$$

相关系数

进而可以使用协方差定义简单相关系数（correlation coefficient）或称皮尔森相关系数（Pearson correlation coefficient）：

$$\rho_{X_1,X_2} = \frac{\mathbb{C}\mathrm{ov}\left(X_1, X_2\right)}{\sqrt{\mathbb{V}\left(X_1\right)\mathbb{V}\left(X_2\right)}}$$

由于

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(X_1, X_2\right) &= \mathbb{E}\left[\left(X_1 - \mathbb{E}\left(X_1\right)\right)\left(X_2 - \mathbb{E}\left(X_2\right)\right)\right] \\
&\le \mathbb{E}\left|\left(X_1 - \mathbb{E}\left(X_1\right)\right)\left(X_2 - \mathbb{E}\left(X_2\right)\right)\right| \\
&\le \sqrt{\mathbb{E}\left|\left(X_1 - \mathbb{E}\left(X_1\right)\right)\right|^2 \mathbb{E}\left|X_2 - \mathbb{E}\left(X_2\right)\right|^2} \\
&= \sqrt{\mathbb{V}\left(X_1\right)\mathbb{V}\left(X_2\right)}
\end{aligned}$$

可知$-1 \le \rho_{X_1,X_2} \le 1$。

相关系数

- 如果$\rho_{Y,Z}=\pm 1$，那么$P\left(Y=c_1Z+c_2\right)=1, c_1\neq 0$，此时$X_1$和$X_2$之间存在完美的线性关系；
- 如果$\rho_{Y,Z}>0$，我们称随机变量Y和Z正相关，反之称为负相关;
- 如果$\rho_{Y,Z}=0$，我们称随机变量Y和Z不相关（uncorrelated），记为$X_1\perp X_2$。

相关系数

注意这里的相关系数实际上只度量了随机变量之间的线性相关性。相关系数等于0并不意味着两个随机变量没有非线性的相关性

简单相关系数与非线性相关

如果随机变量$Y = Z^2$，$Z \sim N(0,1)$，那么：

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Z,Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\ &= \mathbb{E}Z^3 \\ &= 0\end{aligned}$$

两者相关系数为0，然而显然两者存在着非线性的函数关系。

协方差

联合密度函数

如果a,b为任意实数，Y和Z为一元随机变量，那么：

$$\begin{aligned}\mathbb{V}\left(aX_1+bX_2\right) &= \mathbb{E}\left(aX_1+bX_2\right)^2-\left[a\mathbb{E}\left(X_1\right)+b\mathbb{E}\left(X_2\right)\right]^2 \\ &= \mathbb{E}\left(a^2X_1^2+b^2X_2^2+2abX_1X_2\right) \\ &\quad -\left[a^2\left(\mathbb{E}\left(X_1\right)\right)^2+b^2\left(\mathbb{E}\left(X_2\right)\right)^2+2ab\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\right] \\ &= a^2\mathbb{V}\left(X_1\right)+b^2\mathbb{V}\left(X_2\right)+2ab\mathbb{C}\mathrm{ov}\left(X_1,X_2\right)\end{aligned}$$

如果$X_1\perp X_2$，那么

$$\mathbb{V}\left(aX_1+bX_2\right)=a^2\mathbb{V}\left(X_1\right)+b^2\mathbb{V}\left(X_2\right)$$

协方差矩阵

如果对于一个随机向量：$X = [X_1, X_2, ..., X_n]'$，我们可以定义矩阵：

$$\begin{aligned}\mathbb{V}(X) &= [\mathbb{C}\mathrm{ov}(X_i, X_j)] \\ &= \begin{bmatrix} \mathbb{V}(X_1) & \mathbb{C}\mathrm{ov}(X_1, X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_1, X_n) \\ \mathbb{C}\mathrm{ov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(X_n, X_1) & \mathbb{C}\mathrm{ov}(X_n, X_2) & \cdots & \mathbb{V}(X_n) \end{bmatrix}\end{aligned}$$

易知协方差矩阵为实对称矩阵。

协方差矩阵的计算

- 根据协方差矩阵的定义，协方差矩阵可以如下计算：

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mathbb{E}(X)][X - \mathbb{E}(X)]'\right)$$

注意X为列向量，从而：

$$X - \mathbb{E}(X) = \begin{bmatrix} X_1 - \mathbb{E}(X_1) \\ \vdots \\ X_n - \mathbb{E}(X_n) \end{bmatrix}$$

从而：$\mathbb{E}[X - \mathbb{E}(X)][X - \mathbb{E}(X)]' =$

$$\mathbb{E}\begin{bmatrix} (X_1 - \mathbb{E}(X_1))^2 & (X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_1 - \mathbb{E}(X_1))(X_n - \mathbb{E}(X_n)) \\ (X_2 - \mathbb{E}(X_2))(X_1 - \mathbb{E}(X_1)) & (X_2 - \mathbb{E}(X_2))^2 & \cdots & (X_2 - \mathbb{E}(X_2))(X_n - \mathbb{E}(X_n)) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mathbb{E}(X_n))(X_1 - \mathbb{E}(X_1)) & (X_n - \mathbb{E}(X_n))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_n - \mathbb{E}(X_n))^2 \end{bmatrix}$$

协方差矩阵的计算

根据定义，有：

$$
\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}\left[(X-\mathbb{E}X)(X-\mathbb{E}X)'\right] \\
&= \mathbb{E}\left[XX' - X\mathbb{E}(X') - \mathbb{E}(X)X' + \mathbb{E}(X)\mathbb{E}(X')\right] \\
&= \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')
\end{aligned}
$$

协方差矩阵的性质

- 根据协方差矩阵的定义，对于任意的n维向量c，我们有：

$$
\begin{aligned}
c'\mathbb{V}(X)c &= c'\left[\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)'\right]c \\
&= \mathbb{E}\left[c'(X-\mathbb{E}X)(X-\mathbb{E}X)'c\right] \\
&= \mathbb{E}\left\{\left[c'(X-\mathbb{E}X)\right]\left[c'(X-\mathbb{E}X)\right]'\right\} \\
&= \mathbb{E}\left[\left(\left[c'(X-\mathbb{E}X)\right]\right)^2\right] \geq 0
\end{aligned}
$$

因而协方差矩阵是一个半正定矩阵，我们记为$\mathbb{V}(X) \succeq 0$。

协方差矩阵的性质

- 当X的分量之间存在完美的线性关系时，即存在一个向量a使得$a'X=\sum_{i=1}^{n}a_iX_i=0$以概率1成立（$P\left(a'X=0\right)=1$），从而自然有$\mathbb{E}\left(a'X\right)=0$，那么

$$a'\mathbb{V}\left(X\right)a=\mathbb{E}\left[\left(\left[a'\left(X-\mathbb{E}X\right)\right]\right)^2\right]=0$$

  此时等号成立。
- 否则，如果X的分量之间不存在完美的线性关系，那么$\mathbb{V}\left(X\right)$为正定矩阵，记为$\mathbb{V}\left(X\right)\succ 0$。

协方差矩阵的性质

线性变换的协方差矩阵

对于一个实数矩阵$A_{h\times n}$，以及h维向量$b=[b_1,\dots,b_h]'$，有

$$\mathbb{V}(AX+b)=A\mathbb{V}(X)A'$$

Proof.

令$A_{h\times n}=[a_1,a_2,\dots,a_h]$以及$b=[b_1,\dots,b_h]'$，有:

$$\begin{aligned}\mathbb{V}(AX+b)&=\mathbb{E}\left[(AX+b-\mathbb{E}(AX+b))(AX+b-\mathbb{E}(AX+b))'\right]\\&=\mathbb{E}\left[(AX-\mathbb{E}(AX))(AX-\mathbb{E}(AX))'\right]\\&=\mathbb{E}\left[(AX-A\mathbb{E}(X))(X'A'-\mathbb{E}(X')A')\right]\\&=\mathbb{E}\left[AXX'A'-AX\mathbb{E}(X')A'-A\mathbb{E}(X)X'A'+A\mathbb{E}(X)\mathbb{E}(X')A'\right]\\&=A\left[\mathbb{E}(XX')-\mathbb{E}(X)\mathbb{E}(X')\right]A'\\&=A\mathbb{V}(X)A'\end{aligned}$$

随机变量的独立性

随机变量的独立性

如果$[X_1, ..., X_n]'$是定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量，如果对于任意的Borel集$\{B_i, 1 \leq i \leq n\}$，有：

$$\mathscr{P}\left(\bigcap_{i=1}^{n} (X_i(\omega) \in B_i)\right) = \prod_{i=1}^{n} \mathscr{P}(X_i(\omega) \in B_i)$$

那么我们称随机变量$\{X_i, 1 \leq i \leq n\}$相互独立。

对于两个随机变量$[X_1, X_2]'$，如果相互独立，可以简记为$X_1 \perp\!\!\!\perp X_2$。

随机向量的独立性

对于随机向量，以上定义等价于：

随机向量的独立性

随机向量$[X_1, ..., X_n]'$各分量相互独立的充要条件是其联合分布函数等于边缘分布乘积：

$$F\left(x_1, ...x_n\right) = P\left(X_1 \leq x_1, ..., X_n \leq x_n\right) = \prod_{i=1}^{n} P\left(X_i \leq x_i\right) = \prod_{i=1}^{n} F_{X_i}\left(x_i\right)$$

如果密度（质量）函数存在，那么：

$$f\left(x_1, ..., x_n\right) = \prod_{i=1}^{n} f_{X_i}\left(x_i\right)$$

随机向量的独立性

独立性

若概率质量函数为：

$X_2 \backslash X_1$	2	3	4	5	6	7	8	F_{X_2}	f_{X_2}
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_{X_1}	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_{X_2}$ $\Vert$
f_{X_1}	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_{X_1} =$	1

可见$f_{X_1,X_2} \neq f_{X_1} \cdot f_{X_2}$，所以随机变量$X_1$与$X_2$不独立。

随机向量的独立性

独立性

两个联合分布函数：

$$F_{U,V}(u,v) = \min\{u,v\}$$
$$\tilde{F}_{U,V}(u,v) = u \cdot v$$

其边缘分布都为均匀分布，即$F_U(u) = u, F_V(v) = v$，然而由于：

$$F_{U,V}(u,v) = \min\{u,v\} \neq F_U(u) \cdot F_C(v)$$
$$\tilde{F}_{U,V}(u,v) = u \cdot v = F_U(u) \cdot F_C(v)$$

因而联合分布服从$F_{U,V}(u,v)$的随机变量不是相互独立的，而服从$\tilde{F}_{U,V}(u,v)$的随机变量是相互独立的。

随机变量函数的独立性

随机变量函数的独立性

$[X_1, ..., X_n]'$为一系列相互独立的随机变量，$1 \le n_1 \le n_2 \le \cdots \le n_k = n$，对于函数$f_1, f_2, ...f_k$，随机向量

$$\left[f_1\left(X_1, ...X_{n_1}\right), f_1\left(X_{n_1+1}, ...X_{n_2}\right), \cdots, f_k\left(X_{n_{k-1}+1}, ...X_{n_k}\right)\right]'$$

的分量也是相互独立的。

独立与期望

独立随机变量乘积的期望

如果随机向量$X = [X_1, X_2]'$的分量X_1和X_2相互独立且可积，那么

$$\mathbb{E}\left(X_1 X_2\right) = \mathbb{E}\left(X_1\right) \mathbb{E}\left(X_2\right)$$

独立随机变量和的矩母函数

如果随机向量$X = [X_1, ..., X_n]'$的分量相互独立，常数向量$a = [a_1, ..., a_n]'$，记

$$S_n = a' X = \sum_{i=1}^{n} a_i X_i$$

那么矩母函数满足

$$M_{S_n}\left(t\right) = \prod_{i=1}^{n} M_{X_i}\left(a_i t\right)$$

独立于不相关

独立与不相关

如果X_1和X_2相互独立且方差存在，那么$\mathbb{C}\mathrm{ov}\left(X_1, X_2\right) = 0$。

不相关与独立

然而反过来，不相关并不意味着独立。

- 比如，$Y = Z^2, Z \sim \mathrm{N}\left(0, 1\right)$
- Y和Z之间不相关，但是$Y = Z^2$存在完美的函数关系，自然不会是独立的。
- 实际上，如果$Y \perp\!\!\!\perp Z$，那么Y和Z的任意函数$g\left(Y\right)$和$h\left(Z\right)$之间都应该是独立的，即$g\left(Y\right) \perp\!\!\!\perp h\left(Z\right)$从而$\mathbb{C}\mathrm{ov}\left(g\left(Y\right), h\left(Z\right)\right) = 0$
- 即如果$Y \perp\!\!\!\perp Z$，那么Y和Z的任意函数之间都应该是不相关的。

条件期望

- 令(Y, X)为一个二元随机向量，如何使用随机变量X预测随机变量Y？
- 在统计中，我们把这类问题成为回归（regression）。
- 如果我们观察到了随机变量X的值，那么X的何种函数形式可以更好的预测Y呢？
- 比较常见的做法是最小化均方误差（mean squared error）：

$$\min_{h \in L^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

其中

$$L^2 = \left\{ h | h : \mathbb{R} \to \mathbb{R}, \mathbb{E}\left[(h(X))^2 \right] < \infty \right\}$$

条件期望

- 定义误差项$\epsilon = Y - h_0(X)$，对于随机变量X的任意函数$g(X)$，我们有：

$$\mathbb{E}[\epsilon \cdot g(X)] = 0$$

- 如果令$g(X) = 1$，那么我们有

$$\mathbb{E}[\epsilon \cdot g(X)] = \mathbb{E}[\epsilon] = \mathbb{E}[Y - h_0(X)] = 0$$

因而$\mathbb{E}(Y) = \mathbb{E}(h_0(X))$。

  - 从而$\mathbb{E}[\epsilon \cdot g(X)] = 0$意味着

$$\mathbb{C}\mathrm{ov}(\epsilon, g(X)) = \mathbb{E}[\epsilon \cdot g(X)] - \mathbb{E}(\epsilon)\mathbb{E}[g(X)] = 0$$

即ϵ与X的任意函数都不相关。

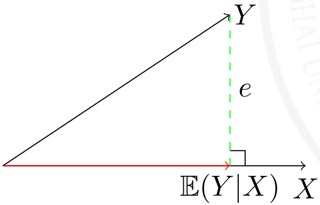
条件期望

- 通过反证法证明，如果存在$g(X)$使得$\mathbb{E}[\epsilon \cdot g(X)] \neq 0$，那么我们令

$$h(X) = h_0(X) + \frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]} g(X)$$

根据这一构造，有：$\mathbb{E}\left[(Y - h(X))^2\right] < \mathbb{E}\left[(Y - h_0(X))^2\right]$，与$h_0(X)$最小化了均方误差矛盾。

- 我们称$h(X)$为Y在X上的正交投影（orthogonal projection）。



条件期望

- 我们知道，

$$\mathbb{E}\left(Y\right) = \arg\min_{c\in\mathbb{R}} \left\{\mathbb{E}\left(Y-c\right)^2\right\}$$

- 仿照上式，我们可以定义随机变量Y给定X的条件期望（conditional expectation）：

$$\mathbb{E}\left(Y|X\right) = h_0\left(X\right) = \arg\min_{h\in\mathscr{L}^2} \left\{\mathbb{E}\left[\left(Y-h\left(X\right)\right)^2\right]\right\}$$

因而随机变量Y给定X的条件期望$\mathbb{E}\left(Y|X\right)$是一个关于$X$的函数。

条件期望与期望

- 期望可以看做是没有任何其他信息时的最优预测，即只用常数对Y进行预测，是条件期望的特例：

$$\mathbb{E}(Y|c) = c^* = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y - c)^2 \right] \right\}$$

- 这也就意味着：
  - 期望本身是对一个随机变量的最优预测
  - 具体的一个实现与期望之间的差异为误差项
  - 比如：
    - 如果全国所有人的平均体重为60公斤，那么随机从人群中选取一个人，对其体重的最优预测为60公斤
    - 单独每个人的体重与60公斤之间的差距为误差项

条件期望：离散情形

现在，进一步，如果我们能看到一个变量$D \in \{0,1\}$，那么：

$$\mathbb{E}\left(Y|D\right) = \arg\min_{h \in \mathbb{H}} \left\{\mathbb{E}\left[\left(Y - h\left(D\right)\right)^2\right]\right\} = \begin{cases} \mathbb{E}\left(Y|D=1\right) = h\left(1\right) & D=1 \\ \mathbb{E}\left(Y|D=0\right) = h\left(0\right) & D=0 \end{cases}$$

- 比如，现在我们可以观察到性别（$D=1$代表男性），全国所有男性的平均体重为70公斤，所有女性平均体重为50公斤
- 那么：

$$\begin{cases} \mathbb{E}\left(Y|D=1\right) = 70 \\ \mathbb{E}\left(Y|D=0\right) = 50 \end{cases}$$

即条件期望为分组平均

条件期望：连续情形

或者，如果我们能看到一个变量X为连续型变量，那么

$$\mathbb{E}(Y|X) = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

为一个未知的函数：

- 比如，如果我们现在可以观察到身高（X）
- 可以假想如果有无数个身高一样的人的平均体重，如：

$$\mathbb{E}(Y|X = 170)$$

即为条件期望。

条件期望的性质

条件期望的性质

对于任意的可测函数$g(X)$，条件期望有如下性质：

1. $\mathbb{E}[g(X)|X]=g(X)$;
2. $\mathbb{E}[(Y-\mathbb{E}(Y|X))\cdot g(X)]=0$;
3. $\mathbb{E}[\mathbb{E}(Y|X)]=\mathbb{E}(Y)$，$\mathbb{E}[Y-\mathbb{E}(Y|X)]=0$;
4. $\mathbb{E}[(g(X)\cdot Y)|X]=g(X)\cdot\mathbb{E}(Y|X)$;
5. $\mathbb{E}(aY_1+bY_2|X)=a\mathbb{E}(Y_1|X)+b\mathbb{E}(Y_2|X)$。

条件期望的性质

银行到达人数

假设每天到达银行的人数服从泊松分布$N \sim P(\lambda)$，而每个到达银行的人，办理外汇业务的概率为p。那么给定到达人数N，办理外汇业务的人数M服从二项分布，即$M|N \sim Bi(N,p), N \sim P(\lambda)$。那么每天来银行办理外汇业务的人数的期望：

$$\mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = p\mathbb{E}(N) = p\lambda$$

均值独立

- 注意到如果我们没有任何信息，因而只能用常数c去预测Y，那么以上最小化问题：

$$\mathbb{E}(Y|c) = c^* = \arg\min_{h \in \mathscr{L}^2} \left\{ \mathbb{E}\left[(Y-c)^2 \right] \right\}$$

对以上最优化问题求解，即：

$$\frac{\partial \mathbb{E}\left[(Y-c)^2 \right]}{\partial c} = \mathbb{E}\left[\frac{\partial (Y-c)^2}{\partial c} \right] = 0$$

从而得到：$\mathbb{E}(Y|c) = \mathbb{E}(Y)$

- 即如果我们没有任何其他随机变量的信息，只能用常数预测Y，那么我们将得到Y的期望。
- 如果有其他随机变量X，但是$\mathbb{E}(Y|X) = \mathbb{E}(Y)$，那么$X$对$Y$的均值没有预测能力，因而我们称$Y$对$X$是均值独立（mean independence）的。

均值独立

- 如果随机变量Y对X是均值独立的，那么：

$$\begin{aligned}\mathbb{C}\text{ov}\left(X,Y\right) &= \mathbb{E}\left(XY\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(XY|X\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(X\mathbb{E}\left(Y|X\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(X\mathbb{E}\left(Y\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)=0\end{aligned}$$

  因而随机变量Y和X必然是不相关的。反之则不成立，不相关并不一定意味着均值独立。
- 实际上，可以证明，$\mathbb{C}\text{ov}\left(g\left(X\right),Y\right)=0$，即$Y$对$X$是均值独立的意味着$Y$与$X$的任意函数都不相关。
- 反过来不一定正确，即$\mathbb{C}\text{ov}\left(g\left(Y\right),X\right)=0$不一定成立（why?）

条件方差

- 相应的，我们还可以定义条件方差

$$\mathbb{V}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2 |X\right]$$

- 根据条件期望的性质：

$$\begin{aligned}
\mathbb{V}(Y|X) &= \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2 |X\right] \\
&= \mathbb{E}\left\{\left[Y^2 + [\mathbb{E}(Y|X)]^2 - 2Y\mathbb{E}(Y|X)\right] |X\right\} \\
&= \mathbb{E}(Y^2|X) + \mathbb{E}\left\{[\mathbb{E}(Y|X)]^2 |X\right\} - 2\mathbb{E}[Y\mathbb{E}(Y|X)|X] \\
&= \mathbb{E}(Y^2|X) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}(Y|X)\mathbb{E}[Y|X] \\
&= \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2
\end{aligned}$$

其中第4个等号由于$\mathbb{E}(Y|X)$也是X的函数

条件方差与方差

根据条件期望的性质，可以证明：

$$\mathbb{V}(Y)=\mathbb{E}\left[\mathbb{V}(Y|X)\right]+\mathbb{V}\left[\mathbb{E}(Y|X)\right]$$

即条件方差的期望与条件期望的方差之和为方差。

条件方差

银行到达人数的方差

在银行到达人数的例子中，可以计算每天办理外汇业务的人数的方差：

$$\mathbb{V}(M) = \mathbb{V}(\mathbb{E}(M|N)) + \mathbb{E}(\mathbb{V}(M|N))$$

其中$\mathbb{E}(M|N) = Np$，因而

$$\mathbb{V}[\mathbb{E}(M|N)] = \mathbb{V}(Np) = p^2\mathbb{V}(N) = p^2\lambda$$

而$\mathbb{V}(M|N) = Np(1-p)$，从而

$$\mathbb{E}(\mathbb{V}(M|N)) = \mathbb{E}(Np(1-p)) = \lambda p(1-p)$$

从而

$$\mathbb{V}(M) = p^2\lambda + \lambda p - \lambda p^2 = \lambda p$$

条件分布

- 如果对于随机向量(X,Y)，我们取$1_A(x)=1\ if\ x\in A\ else\ =0$，这是一个随机变量$X$的函数，因而

$$\mathbb{E}(Y\cdot 1_A(X))=\mathbb{E}[\mathbb{E}(Y|X)\cdot 1_A(X)]=\mathbb{E}[h_0(X)\cdot 1_A(X)]$$

- 若X,Y是离散型随机变量，那么我们令$A=\{X=x_i\}$有：

$$\mathbb{E}(Y\cdot 1\{X=x_i\})=h_0(x_i)\cdot P(X=x_i)$$

从而：

$$\begin{aligned}\mathbb{E}(Y|X=x_i)=h_0(x_i)&=\frac{\mathbb{E}(Y\cdot 1\{X=x_i\})}{P(X=x_i)}\\&=\frac{\sum_{k=0}^{\infty}[y_k\cdot P(Y=y_k,X=x_i)]}{P(X=x_i)}\\&=\sum_{k=0}^{\infty}y_k\cdot\frac{P(Y=y_k,X=x_i)}{P(X=x_i)}\end{aligned}$$

条件分布

- 对于连续型随机向量(X,Y)，可以证明：

$$\mathbb{E}(Y|X=x)=h_0(x)=\frac{\int_{\mathbb{R}} yf(x,y)\,dy}{f_X(x)}=\int_{\mathbb{R}} y\frac{f(x,y)}{f_X(x)}dy$$

  - 由于$\mathbb{E}[Y-\mathbb{E}(Y|X)]=0$，从而：

$$\int_{\mathbb{R}}[y-h_0(x)]f(x,y)\,dy=0$$

  - 固定x，那么以上条件意味着：

$$\int_{\mathbb{R}} yf(x,y)\,dy=h_0(x)\int_{\mathbb{R}} f(x,y)\,dy=h_0(x)f_X(x)$$

条件分布

条件密度函数

对于离散型随机变量，定义

$$f_{Y|X}(y|x) = \frac{P(Y=y, X=x)}{P(X=x)} = \frac{P(Y=y, X=x)}{\sum_y P(Y=y, X=x)}$$

对于连续型随机变量，定义

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy}$$

我们把$f_{Y|X}(y|x)$定义为条件密度函数（conditional denstiy function）。

条件密度函数

- 条件期望可以写为条件密度函数的积分

$$\mathbb{E}(Y|X) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\,dy$$

- 注意：对于离散型随机变量：

$$\sum_y f_{Y|X}(y|x) = \sum_y \frac{P(Y=y, X=x)}{\sum_y P(Y=y, X=x)} = 1$$

而对于连续型随机变量：

$$\int_{\mathbb{R}} f_{Y|X}(y|x)\,dy = \int_{\mathbb{R}} \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy}dy = 1$$

因而条件密度函数也是密度函数。

独立与均值独立

- 如果随机变量X和Y是独立的，那么：

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)\cdot f_Y(y)}{f_X(x)} = f_Y(y)$$

即两个随机变量独立的充要条件是$f_{Y|X} = f_Y$。

- 在独立的条件下：

$$\mathbb{E}(Y|X) = \int_{\mathbb{R}} y\cdot f_{Y|X}(y|x)\,dy = \int_{\mathbb{R}} y\cdot f_Y(y)\,dy = \mathbb{E}(Y)$$

因而如果随机变量X和Y是独立的，那么其一定是均值独立的，反之则不成立。

- 独立、均值独立、不相关三者的强弱关系如下：

$$X \perp\!\!\!\perp Y \underset{\not\Leftarrow}{\overset{\Rightarrow}{}} \mathbb{E}(Y|X) = \mathbb{E}(Y) \underset{\not\Leftarrow}{\overset{\Rightarrow}{}} X \perp Y$$

其中$X \perp\!\!\!\perp Y$代表X与Y独立，$X \perp Y$代表X与Y不相关。

条件密度函数

四面骰子

四面骰子的例子中，其条件密度可以如下计算：

$Z\backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$
$f_{Y\|Z}(y\|Z=1)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	0	0	
$f_{Y\|Z}(y\|Z=2)$	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0	

条件密度函数

二元正态分布

对于联合正态

$$(X,Y)' \sim N\left(\left(\begin{array}{c}\mu_X\\ \mu_Y\end{array}\right),\left(\begin{array}{cc}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{array}\right)\right)$$

密度函数：

$$f_{X,Y}(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}}\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}$$

其中$-1<\rho<1$为X,Y的相关系数。

条件密度函数

四面骰子

其边际密度函数为：

$$
\begin{aligned}
f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \\
&\cdot \int_{\mathbb{R}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(1-\rho^2)(x-\mu_X)^2}{\sigma_X^2}+\left(\frac{y-\mu_Y}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right]\right\}dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}
\end{aligned}
$$

或者$X \sim N\left(\mu_X, \sigma_X^2\right)$

条件密度函数

四面骰子

其条件密度函数为:

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}
\end{aligned}
$$

或者$Y|X \sim N\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right)$，也服从正态分布，进而:

- 条件期望$\mathbb{E}(Y|X=x) = \mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$
- 条件方差$\mathbb{V}(Y|X) = \sigma_Y^2(1-\rho^2)$。

贝叶斯公式

- 使用条件密度函数的定义，我们还可以得到随机变量的贝叶斯公式。
- 由于：$f(x,y) = f_X(x) \cdot f_{Y|X}(y|x) = f_Y(y) \cdot f_{X|Y}(x|y)$从而条件密度：

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{\int_{\mathbb{R}} f(x,y)\, dy}$$
$$= \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y) \cdot f_Y(y)\, dy}$$

以上方程即随机变量的贝叶斯公式，在贝叶斯统计中有大量的应用。

条件分布与统计模型

- 统计模型通常研究几个随机变量的联合分布，如$F(X,Y)$
- 经常我们不需要对联合分布进行建模，而是通过条件分布对联合分布进行建模。由于：
$$f(x,y) = f_X(x)\cdot f_{Y|X}(y|x)$$
因而我们可以通过对X的边缘分布以及$Y|X$的条件分布对X,Y的联合分布进行建模。
- 比如，分层模型（hierarchical model）就通过分层次的假设条件分布对数据的分布进行建模。
- 有时X的边缘分布不是关注的核心问题，甚至可能只对$Y|X$的边缘分布进行建模。

高斯混合模型

高斯混合模型

如果我们关注某一项疾病指标X，该指标对于患者和健康人群具有不同的分布。记$D=1$为患者，$D=0$为健康人群，记患者该项指标为X_1，健康人群该项指标为X_0，假设：

$$\begin{cases} X_1 \sim N\left(\mu_1, \sigma_1^2\right) \\ X_0 \sim N\left(\mu_0, \sigma_0^2\right) \end{cases}$$

即分别假设了患者和健康人群该项指标的分布，那么观察到的指标：$X = DX_1 + (1-D)X_0$。该模型可以写为：

$$\begin{cases} X|D=1 \sim N\left(\mu_1, \sigma_1^2\right) \\ X|D=0 \sim N\left(\mu_0, \sigma_0^2\right) \\ D \sim Ber\left(p\right) \end{cases}$$

高斯混合模型

高斯混合模型

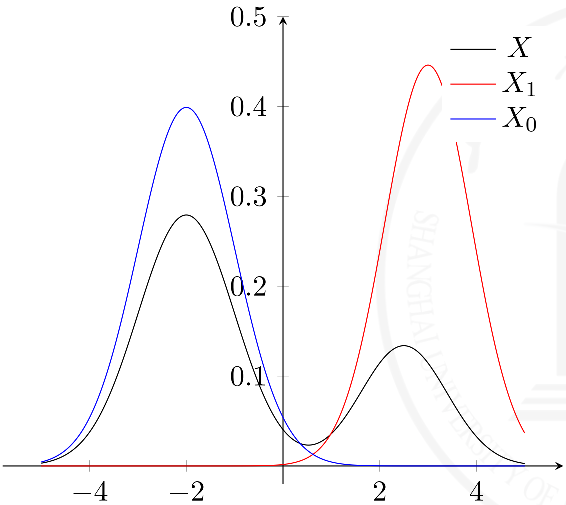
为了得到X的密度函数，注意到：

$$
\begin{aligned}
F_X(x) = P(X \le x) = \mathbb{E}[1(X \le x)] = \mathbb{E}\{\mathbb{E}[1(X \le x)|D]\} \\
= \mathbb{E}[1(X \le x)|D=1] \cdot P(D=1) \\
+ \mathbb{E}[1(X \le x)|D=0] \cdot P(D=0) \\
= \Phi\left(\frac{x-\mu_1}{\sigma_1}\right) \cdot p + \Phi\left(\frac{x-\mu_0}{\sigma_0}\right) \cdot (1-p)
\end{aligned}
$$

从而：

$$
\begin{aligned}
f_X(x) &= p\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-p)\frac{1}{\sigma_0}\phi\left(\frac{x-\mu_0}{\sigma_0}\right) \\
&= pf_{X_1}(x) + (1-p)f_{X_0}(x)
\end{aligned}
$$

高斯混合模型



高斯混合模型

高斯混合模型

我们可以使用条件期望计算X的期望：

$$
\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}\left[\mathbb{E}(X|D)\right] = \mathbb{E}\left\{\mathbb{E}\left[DX_1 + (1-D)X_0|D\right]\right\} \\
&= \mathbb{E}\left\{D\mathbb{E}(X_1|D) + (1-D)\mathbb{E}(X_0|D)\right\} \\
&= \mathbb{E}\left\{D\mu_1 + (1-D)\mu_0\right\} \\
&= \mu_1\mathbb{E}(D) + \mu_0\mathbb{E}(1-D) \\
&= p\mu_1 + (1-p)\mu_0
\end{aligned}
$$

高斯混合模型

高斯混合模型

此外，如果我们观察到了X，也可以使用贝叶斯公式计算其患病的概率：

$$
\begin{aligned}
f_{D|X}(d=1|x) &= \frac{f_{X|D}(x|d=1)\, f_D(d=1)}{\int_{\mathbb{R}} f_{X|D}\left(x|\tilde{d}\right) f_D(d)\, d\tilde{d}} \\
&= \frac{\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) p}{\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) p + \frac{1}{\sigma_0}\phi\left(\frac{x-\mu_0}{\sigma_0}\right)(1-p)}
\end{aligned}
$$

迭代期望公式

- 条件期望可以很方便的扩充到多个X的情形，比如$\mathbb{E}\left(Y|X_1,X_2\right)$可以定义为:

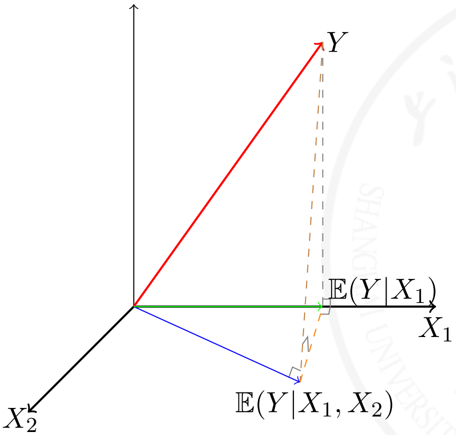
$$\mathbb{E}\left(Y|X_1,X_2\right)=h_0\left(X_1,X_2\right)=\arg\min_{h\in\mathscr{L}^2}\left\{\mathbb{E}\left[\left(Y-h\left(X_1,X_2\right)\right)^2\right]\right\}$$

- 条件期望有如下性质（**迭代期望公式，law of iterated expectation**）：

$$\mathbb{E}\left[\mathbb{E}\left(Y|X_1,X_2\right)|X_1\right]=\mathbb{E}\left(Y|X_1\right)$$

即如果我们对随机变量Y，先在大的空间上投影，再在这个大的空间上的一个小的子空间上进行投影，与直接在这个小的空间上进行投影是相等的。

迭代期望公式



条件期望

条件期望与$\sigma-$代数

在四面骰子的例子中，随机变量Z可能取值为：$\{1,2,3,4\}$，因而：

$$\sigma\langle Z\rangle = \sigma\left\langle Z^{-1}(A) : A\in\mathcal{B}\right\rangle = \\ \sigma < \{(1,1),(2,1),(3,1),(4,1),(1,2),(1,3),(1,4)\}, \\ \{(2,2),(2,3),(2,4),(3,2),(4,2)\}, \\ \{(3,3),(3,4),(4,3)\},\{(4,4)\} >$$

例如，如果我们只知道$Z=3$，我们知道实际发生的情况应该是$\{(3,3),(3,4),(4,3)\}$中的某一种。因而如果给定$Z=3$，我们把之前的16种情况降低到了3种情况。如果我们对Y，即两个骰子的和感兴趣，如果我们没有任何信息，那么我们对Y的最优预测应该是$\mathbb{E}(Y)=5$。而如果我们观察到了$Z=3$，那么此时最优预测应该为$\mathbb{E}(Y|Z=3)=\frac{6+7+7}{3}=\frac{20}{3}$。

条件期望

条件期望与$\sigma-$代数

- 在上例中，Z总共有4种可能的取值，在每种Z的可能取值的情况下，都可以把16种情况降低为更少的情况，因而增大了信息量。
- 而如果我们使用随机变量Y，Y共有7种可能的取值，给定Y也会增大我们的信息量。
- 而如果给定(Z, Y)两个随机变量，可以更加细分为10种情况，我们可以得到

$$\sigma\langle Z\rangle \subset \sigma\langle Z, Y\rangle, \sigma\langle Y\rangle \subset \sigma\langle Z, Y\rangle$$

即两个随机变量提供了比单独一个随机变量更多的信息。

  - 例如，如果我们不仅仅观察到$Z=3$，还观察到$Y=7$，那么我们此时知道，实际发生的情况应该是$\{(3,4),(4,3)\}$两种情况下的一种，比只观察到Z时更加准确。

条件期望定义

因而我们通常把条件期望的概念推广到$\sigma-$代数上。对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，我们可以对$\mathscr{F}$的一个子$\sigma-$代数$\mathscr{G} \subset \mathscr{F}$定义条件期望如下：

条件期望

对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，$\mathcal{I} \subset \mathscr{F}$为一个$\sigma-$代数，对于随机变量$Y$满足：$\mathbb{E}\left(|Y|\right) < \infty$，如果对于任意的$A \in \mathcal{I}$，随机变量$H$满足：

$$\mathbb{E}\left(Y \cdot 1_A\right) = \mathbb{E}\left(H \cdot 1_A\right)$$

那么我们称H为给定$\mathcal{I}$随机变量Y的条件期望，记为$\mathbb{E}\left(Y|\mathcal{I}\right)$。令$B \in \mathscr{F}$，定义$\mathscr{P}\left(B|\mathcal{I}\right) = \mathbb{E}\left(1_B|\mathcal{I}\right)$为条件概率。

迭代期望公式

- 注意以上定义的$\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma\langle X\rangle)$。
- 特别的，令$\mathcal{I} = \{\emptyset, \Omega\}$，$\mathbb{E}(Y|\{\emptyset, \Omega\}) = \mathbb{E}(Y)$，即信息量最小的条件期望即为期望本身。
- 而以上的迭代期望公式也可以相应推广，即如果$\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathscr{F}$，那么：

$$\mathbb{E}(Y|\mathcal{I}_1) = \mathbb{E}\{[\mathbb{E}(Y|\mathcal{I}_2)]|\mathcal{I}_1\}$$

即先在大的信息集上做投影，再将其投影到小的信息集上，等价于直接投影在小的信息集上。

作业

- 3.2， 3.4， 3.6， 3.8， 3.9, 3.10, 3.15, 3.18